REGISTRATION OF MULTIMODAL DATA FOR ESTIMATING THE PARAMETERS OF AN ARTICULATORY MODEL

M. Aron, A. Toutios, M.-O. Berger, E. Kerrien, B. Wrobel-Dautcourt, Y. Laprie

LORIA/ CNRS/ INRIA Nancy Grand-Est
BP 101, 54602 Villers les Nancy, France

ABSTRACT

Being able to animate a speech production model with articulatory data would open applications in many domains. In this paper, we first consider the problem of acquiring articulatory data from non invasive image and sensor modalities: dynamic ultrasound (US) images, stereovision 3D data, electromagnetic sensors and MRI. We here especially focus on automatic registration methods which enable the fusion of the articulatory features in a common frame. We then derive articulatory parameters by fitting these features with Maeda's model. To our knowledge, it is the first attempt to derive articulatory parameters from features automatically extracted and registered between the modalities. Results prove the soundness of the approach and the reliability of the fused articulatory data.

Index Terms— speech analysis, multimodal registration

1. INTRODUCTION

The animation of a speech production model with articulatory data would open applications in many domains. Learners of a foreign language, for instance, would get articulatory feedback to discover the articulation of sounds that do not exist in their mother tongue.

There are however two obstacles. The first is the difficulty of acquiring articulatory and especially dynamic articulatory data. Even if recent advances in MRI [1] have been realized the technique is expensive and not easily available. In the framework of the European ASPI project we have developed an acquisition system combining ultrasound (US) imaging, stereovision and electromagnetic (EM) sensors. It offers the possibility of capturing, with a good time resolution, a continuous contour of the tongue in the mouth cavity via ultrasound imaging (at a frequency of 66Hz) and 3D information of the speaker's face (at a frequency of 120Hz). Additionally, the ASPI system utilizes two EM sensors glued onto the tongue tip and tongue blade (acquired at 40 Hz) to get the tongue tip position often hidden by the jaw bone and air in US images. The acquisition of dynamic data without too many constraints imposed on the speaker and at a low cost are the main advantages of this system.

The second obstacle is to derive usable articulatory information, here parameters to animate an articulatory model. We thus need to fuse these different sources of articulatory information with some static MRI images in order to get a 3D reference geometry of the vocal tract, and more importantly, to adjust parameters of a flexible articulatory model. Here, we make the assumption that we have an existing flexible articulatory model at our disposal. This means that this model, that of Maeda [2] in our case, is expected to reasonably well account for a new speaker.

The authors acknowledge the financial support of the FET programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324.

This work comprises two aspects: (i) fusing the different imaging and articulatory modalities – this mainly consists in registering US images and stereovision data within the MRI images– , (ii) deriving articulatory parameters from these data – this consists in adapting the model to the speaker and in optimizing the set of parameters to describe the measured vocal tract shape.

2. MULTIMODAL REGISTRATION

2.1. Acquisitions and processing of the data

The ASPI system was initially designed to get temporally and spatially aligned acquisitions of US, EM and audio data [3]. It has been extended to acquire synchronized stereovision data using our system described in [4]. This system allows us to obtain the temporal 3D coordinates of points painted on the lips and on the chin (Fig. 2.a).

A corpus of data (20 min and 30 seconds, 81180 US images and 145140 stereo images) was acquired for a speaker, including single vowels (/a/, /i/..), Vowel-Vowel (/ae/,...), Vowel-Consonant-Vowel (/iku/,...) and complete sentences in French.

In order to extract the articulatory features -tongue, lips,...- from this massive amount of data efficiently, we have developed an automatic tracking tool to extract the tongue shape on the US images [3]. Automatic stereovision techniques have been developed to track feature points on the speaker's face and therefore recover the trajectory of these points over the whole video sequence. Thus, the lip parameters (width, height, and protrusion) can be easily computed from these stereo data.

A high-resolution MRI of a neutral expression of the head is also acquired. This allows us to extract static articulory data such as the palate. This volume is also used as a reference to register the image modalities.

2.2. Registration of data

Only some articulatory parameters are visible in the various modalities: the tongue can be extracted from the US image, the palate is only visible in the MRI images, lip protrusion can be computed from stereovision data. We thus have to register these data and to compensate for head motion in order to express all the articulatory features in the same coordinate system.

Direct registration is not possible because no common feature points are visible in the modalities. We thus use extra EM sensors placed on the head and on the US probe to get features which are measured both in the EM coordinate system and in the stereovision system or the US system. This allows us to compute the transformations $T_{mri \leftarrow stereo}$, $T_{mri \leftarrow us}$ and thus to fuse the articulatory features in the MRI frame.

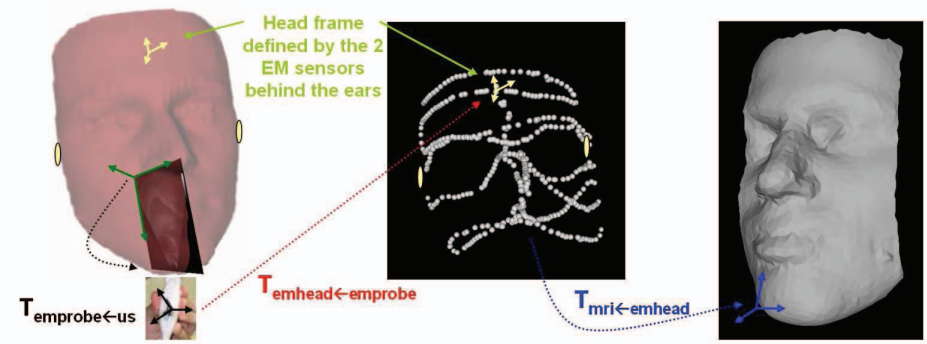


Fig. 1. Intermediate transformations involved in the registration between the US modality and the MRI modality.

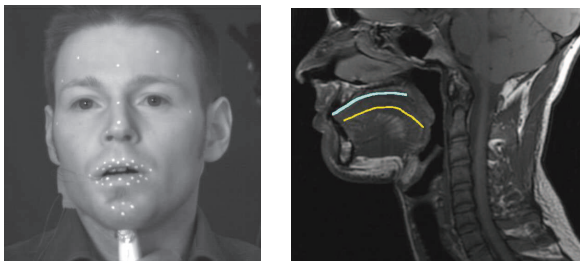
Firstly, the rigid transformation $T_{emprobe \leftarrow us}$ between the US image coordinate system and the EM coordinate system attached to the US probe is computed via the calibration procedure detailed in [3]. In order to compensate for head movements during the acquisitions, data are expressed in the frame defined by two 5 dof (degrees of freedom) sensors glued under the ears of the speaker (Fig. 1), giving rise to the $T_{emhead \leftarrow emprobe}$ transformation.

The rigid transformation $T_{mri \leftarrow emhead}$ is calculated thanks to an EM scan of the face made with a pointer at the beginning of the acquisition session on a neutral expression. These scanned points are expressed into the EM head coordinate system, and the Iterative Closest Point (ICP) method [5] is used to compute the rigid transformation $T_{mri \leftarrow emhead}$ between the EM scan and the face extracted from the MRI. The composition of these intermediate rigid transformations gives the complete transformation $T_{mri \leftarrow us}$ between the US and the MRI data (Fig. 1).

$T_{mri \leftarrow stereo}$ is estimated through the use of the ICP algorithm between the set of 3D points computed by stereovision on the neutral expression and the surface of the face extracted from the MRI.

Fig. 2.b exhibits the superimposition of a tongue shape extracted onto the MRI for the vowel /a/.

These fused data are then used for the estimation of the articulatory parameters of the vocal tract.



(a) (b)

Fig. 2. Data acquisition and registration : (a) a stereo image of our speaker painted with markers. (b) US tongue on a /a/ (in yellow) superimposed on MRI using T_{us2mri} . The palate extracted from MRI has been drawn (in blue).

3. DERIVATION OF ARTICULATORY PARAMETERS

Maeda's model [6] describes the vocal tract shape in the form of a weighted sum of 7 linear components. The vocal tract is represented by its midsagittal profile with reference to a semipolar coordinate grid system which consists of a polar part and two linear parts for the buccal and pharyngeal areas (see Fig. 3).

The objective is to find the weights attached to each linear component (articulatory parameters) in order that the articulatory model best approximates data provided by the system above, while preserving naturalness (as implied by the model) and slow changing rate of the corresponding vocal tract shapes. In the rest of this section we propose a solution to this problem and present preliminary results.

3.1. Speaker adaptation of the model

The semipolar grid system was initially defined for a specific female speaker and had to be adjusted to the male speaker of the present study. The necessary adjustments were performed manually by superimposing the grid system on an MRI picture of the speaker, and involved the determination of the mouth and pharynx scale factors, as well as the correction of the contour of the fixed external part of the vocal tract corresponding to the hard palate and the pharyngeal wall. The scale factors were determined to be 1.09 for the mouth and 1.11 for the pharynx. The corrected contour of the external part of the vocal tract can be seen in Fig. 3.

3.2. From ultrasound to tongue variables

Variables used to calculate tongue articulatory parameters were the intersections of the tongue contour with the grid lines. They were derived from the tongue contour provided by the ultrasound images.

The registration of the ultrasound tongue contours with the semipolar grid system was readily achieved via the knowledge of its registration with the MRI vocal tract. Fig. 3 shows the derived intersections. These were then normalized, using statistics included in the original definition of the model.

3.3. From stereovision to jaw and lip variables

Additional variables used in the context of the articulatory model described the jaw position, lip opening, lip protrusion and lip width. Their values can be derived on the basis of the stereovision data.

Positions of stereovision marks in the middle of the upper and lower lip, after projection on carefully selected axes, provided the lip

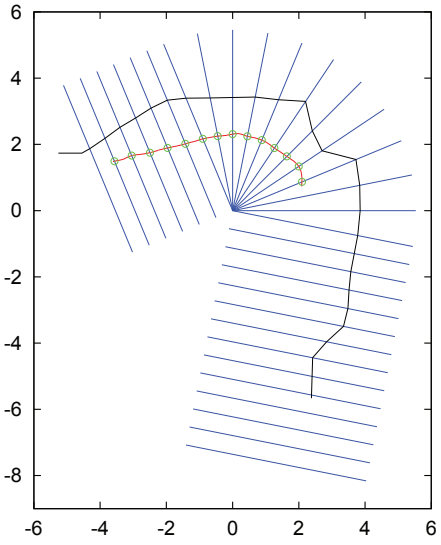


Fig. 3. Semipolar grid system, external vocal tract, ultrasound tongue contour and values at the gridlines (circles). Part of an /a/ utterance.

opening and protrusion variables. Marks on the lip corners provided the lip width variable. A mark on the mandible provided the jaw variable. These variables were z-scored using statistics derived from the whole recording session.

3.4. From variables to parameters

In the model, the variables already described are generated from a set of underlying parameters via a linear relationship:

$$\mathbf{v} = \mathbf{A}\mathbf{p} \tag{1}$$

where $\mathbf{v}$ is a 29-dimensional vector (consisting of the jaw variable, the 3 lip variables and 25 tongue variables); $\mathbf{A}$ is a 29×6 table of real values; and $\mathbf{p}$ is a 6-dimensional vector of parameters (jaw position, tongue body position, tongue shape, apex position, lip opening and lip protrusion). The model's larynx parameter is excluded from our present discussion, since no information on the state of the larynx may be derived from ultrasound or stereovision data.

Given a subset of measured variables, say C, at a given time instant, we can approximate the corresponding parameters by minimizing the quantity:

$$I_s = \sum_{i\in C}\left(v_i - \sum_{j=1}^{6} a_{i,j}p_j\right)^2 \tag{2}$$

where $a_{i,j}$ are elements of matrix $\mathbf{A}$ and p_j elements of vector $\mathbf{p}$. We further introduced the constraints:

$$p_j \in [-3,3],\ j = 1,\ldots,6 \tag{3}$$

which ensure that the vocal tract configurations corresponding to the estimated parameters are realistic. The minimization of Eq. (2) subject to Eq. (3) constitutes a regular constrained quadratic programming problem that is solved using an active space null set method [7].

To ensure the smoothness of the articulatory trajectories over time, we used a regularization technique based on variational calculus [8]. We introduced the following cost function to be minimized over the time interval $[t_s, t_f]$:

$$\begin{aligned} I_d = &\int_{t_s}^{t_f}\sum_{i\in C}\left(v_i(t) - \sum_{j=1}^{6} a_{i,j}p_j(t)\right)^2 dt \\ &+\lambda\int_{t_s}^{t_f}\sum_{j=1}^{6} p_j'(t)^2 dt + \beta\int_{t_s}^{t_f}\sum_{j=1}^{6} p_j(t)^2 dt \end{aligned} \tag{4}$$

The three integrals express respectively: proximity between observed and model generated variables; changing rate of parameters; and total articulatory effort. Constants λ and β were chosen heuristically. The minimization of I_d gives rise to an iterative process that updates the parameter values at each step, until the satisfaction of a convergence criterion [9]. The articulatory parameter vectors found by minimizing Eq. (2) subject to Eq. (3) for the discrete-time samples in the interval $[t_s, t_f]$ were used as a startup solution for this iterative process.

3.5. Results

Fig. 4 shows the estimated trajectories of model parameters for the uttered vowel-vowel sequence /ae/. Fig. 5 shows the evolution of the vocal tract shapes derived on the basis of these trajectories. The sequence /ae/ is presented because the acoustic parameters of these two sounds are not too far from each other. Thus, recovering a relevant articulatory transition for this sequence proves the soundness of the approach.

First, it turned out that the fusion of the different modalities was successful. This is an important conclusion since it means that it is possible to recover relevant global articulatory information from several modalities. Only one static MRI image is necessary to register all the modalities together.

Second, as it can be seen in the series of vocal tract images corresponding to the sequence /ae/, the US tongue contours enabled relevant global shapes of the vocal tract to be recovered. The main articulatory feature, i.e. the location of the main constriction of the vocal tract, was tracked successfully: it moves from the pharyngeal region characteristic of a /a/ vowel to the front part of the vocal tract corresponding to the articulation of /e/.

The articulatory model adaptation focuses on overall geometrical parameters (scale factors of mouth and pharynx cavities, motionless external wall of the vocal tract, i.e. hard palate and pharyngeal wall) but not on its dynamic features, i.e. the modes of deformation of tongue given by the linear components. This explains the poorer results observed in some sequences. Fig. 6 shows four intermediate vocal tract shapes extracted from the sequence /ay/. The last shape could not be interpolated efficiently by the four linear components used to describe the tongue and the optimization thus weakened the front constriction of /y/ in order to get a better overall fitting. Two solutions will be explored in the near future. The first consists in focusing the fitting on the front part of the US tongue contour. This is all the more justified since the back part of the US contour is often less reliable. The second solution consists in adapting the linear components by using some known reference speech utterances and minimizing the difference between the expected articulatory transitions and those recovered.

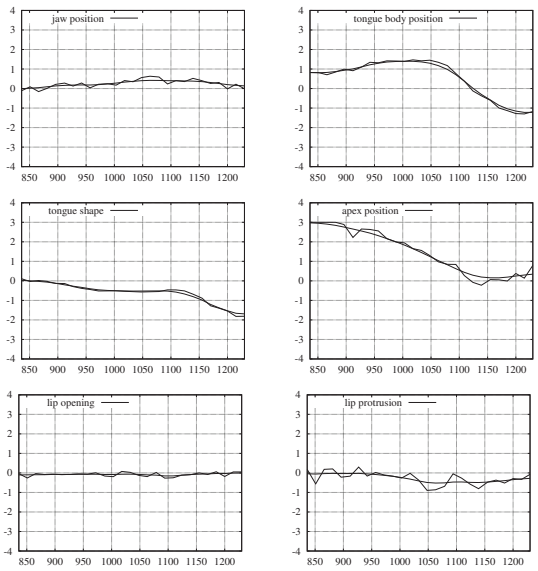


Fig. 4. Articulatory trajectories for the sequence /ae/. In each graph the trajectory obtained by framewise optimization and that obtained by using the variational regularization method (smoothest trajectories) are plotted. The horizontal axis represents time in milliseconds.

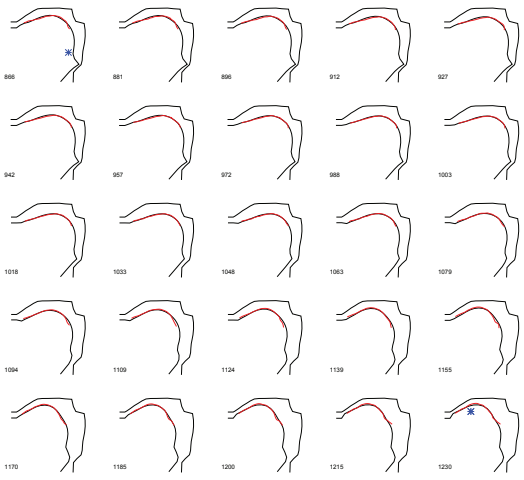


Fig. 5. Registered ultrasound contours and model-derived vocal tract shapes for the vowel-vowel sequence /ae/. Numbers denote time in milliseconds. Asterisks indicate the approximate location of the narrowest constriction in the vocal tract.

Fig. 6. Registered ultrasound contours and model-derived vocal tract shapes for the vowel-vowel sequence /ay/. Numbers denote time in milliseconds.

4. CONCLUSION

The main two contributions of this work are the fusion of several imaging modalities of the vocal tract and the speaker's face, and the recovery of relevant articulatory information from these data. An important point is that only one MRI static image, used as a reference image, is necessary. This approach can thus be envisaged to provide articulatory feedback in speech therapy, to pilot articulatory synthesis or talking heads used in multimedia applications, to train and/or evaluate acoustic to articulatory mapping.

5. REFERENCES

[1] E. Bresch, Y.-C. Kim, K. Nayakand D. Byrd, and S. Narayanan, "Seeing Speech: Capturing Vocal Tract Shaping Using Real-Time Magnetic Resonance Imaging," *IEEE Signal Processing Magazine*, vol. May, pp. 123–132, 2008.

[2] S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires," in *Actes 10èmes Journées d'Etude sur la Parole*, Grenoble, Mai 1979, pp. 152–162.

[3] M. Aron, A. Roussos, M.O. Berger, E. Kerrien, and P. Maragos, "Multimodality Acquisition of Articulatory Data and Processing," in *Proceedings of the European Signal Processing Conference (Eusipco), Lausanne*, 2008.

[4] B. Wrobel-Dautcourt, M. O. Berger, B. Potard, Y. Laprie, and S. Ouni, "A low cost stereovision based system for acquisition of visible articulatory data," in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05)*, Vancouver, 2005, pp. 145–150.

[5] P.J. Besl and N.D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[6] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*, W.J. Hardcastle and A. Marchal, Eds., pp. 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

[7] R. Fletcher, *Practical methods of optimization*, Wiley-Interscience New York, NY, USA, 1987.

[8] M. Bonvalet, *Les principes variationnels*, Masson, 1993.

[9] Y. Laprie and B. Mathieu, "A variational approach for estimating vocal tract shapes from the speech signal," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, May 1998, vol. 2, pp. 929–932.